A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu

B. Sasidhar[#1], P. M. Yohan[*2], Dr. A. Vinaya Babu3, Dr. A. Govardhan[4],

[1] **Associate Professor, Dept. of MCA, CM Engineering College, Secunderabad, Andhra Pradesh, India.**

[2] **Associate Professor, Dept. of MCA, Wesley P.G. College, Secunderabad, Andhra Pradesh, India.**

[3] **Director, Admissions,JNTU, Kukatpally, Hyderabad, Andhra Pradesh, India.**

[4] **Principal, JNTUH College of Engg, Nachupally(Kondagattu), Karimanagar Dt., Andhra Pradesh, India.**

Abstract

In this paper, we present a survey of various approaches for identification of Named Entities (NE) in Indian Languages. First we present various approaches used to recognize NE in Indian languages . Next we critically describe the observations and research related to NER. In the language of English it is observed capitalization is a major clue to identify NEs. Indian languages are resource poor languages and gazetteers available are insufficient. Indian languages are agglutinative in nature the reason being more number of inflectional words.

Keywords: Named Entity, Named Entity Recognition

1. Introduction

A Named entity is any thing about a name. Named Entity recognition is a proper sequence of identification of name and its classification. NER is a main sub task of Information Extraction. Numerous NER applications are found and observed in varied branches of knowledge and science such as Information Extraction, Question-Answering, Machine Translation, Automatic Indexing of documents , Cross-lingual Information retrieval, Text Summarization etc.,.

Telugu is a most popular language in southern part of India. Telugu language occupied 15th position in the world and 2nd position in India. Telugu language belongs to Dravidian family. Telugu is a highly inflectional and agglutinative language. Each word in Telugu is inflected for a very large number of word forms. Telugu is primarily suffixing language, in which several suffixes added to the right. Telugu is a verb final language (in general) and word free order language [1].

A few of the Various Named Entity classes identified in NER are

- Person Name
- Organization Name
- Location Name
- Designation
- Abbreviation
- Brand
- Title person
- Title object
- Number
- Measure
- Term
- Date and Time

2. Approaches on NER

Various approaches used in NER system are Rule based / Handcrafted Approach, Machine Learning / Automated / Statistical approach, and Hybrid Model.

2.1. The Rule based / Handcrafted Approach

2.1.1. List Lookup Approach:

NER system uses gazetteer to classify words. We just have to create a suitable list in the gazetteer. It is simple,

fast and language independent. It is also easy to retarget as we just have to create lists. Only works for lists in the gazetteer. We have to collect and maintain the gazetteer. This approach cannot resolve ambiguity.

2.1.2. Linguistic Approach:

NER system uses some language based rules and other heuristic to classify words. It needs rich and expressive rules and gives good results. It requires an advanced knowledge of grammar and other language related rules. This calls for a thorough knowledge and advanced skills related to the Language under consideration are needed to come up with good rules and heuristic.

2.2. Machine Learning Based Approach / Automated Approach

2.2.1. Hidden Markov Models (HMMs):

It is a generative model. The model assigns a joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets. P(X, Y) =Πi P(Xi , Yi) P(Yi , Yi-1) It uses forward-backward algorithm, Viterbi Algorithm and Estimation-Modification method for modeling. Its basic theory is elegant and easy to understand. Hence it is easier to implement and analyze. In order to define joint probability over observation and label sequence HMM needs to enumerate all possible observation sequence. Hence it makes various assumptions about data like Markovian assumption i.e. current label depends only on the previous label. Also it is not practical to represent multiple overlapping features and long term dependencies. Number of parameter to be evaluated is huge. So it needs a large data set for training.

2.2.2. Maximum Entropy Markov Models (MEMMs):

It is a conditional probabilistic sequence model. It can represent multiple features of a word and can also handle long term dependency. It is based on the principle of maximum entropy which states that the least biased model which considers all know facts is the one which maximizes entropy. Each source state has a exponential model that takes the observation feature as input and output a distribution over possible next state. Output labels are associated with states. It solves the problem of multiple feature representation and long term dependency issue faced by HMM. It has generally increased recall and greater precision than HMM. It has Label Bias Problem. The probability transition leaving any given state must sum to one. So it is biased towards states with lower outgoing transitions. The state with single outgoing state transition will ignore all observations. To handle Label Bias Problem we can change the state-transition structure or we can start with fully connected model and let the training procedure decide a good structure.

2.2.3. Conditional Random Field (CRF):

It is a type of discriminative probabilistic model. It has all the advantage of MEMMs without the label bias problem. CRFs are undirected graphical models (also know as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes. Random field: Let G = (Y, E) be a graph where each vertex YV is a random variable. Suppose P(Yv | all other Y) = P(Yv |neighbors(Yv)), then Y is a random field[2].

Let X = random variable over data sequences to be labeled Y = random variable over corresponding label sequence. “Definition Let G = (V,E) be a graph such that Y = (Yv) v∈V , so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field in case, when conditioned on X, the random variables Yv obey the Markov Property with respect to the graph: P(Yv |X,Yw, w ≠ v) = P(Yv |X,Yw, w ∈ v), where w ∈ v means that w and v are neighbors in G.

2.2.4 Support Vector Machine (SVM):

SVM is one of the famous supervised machine learning algorithms for binary classification in all various data set and it gives the best results where the data set is a few, and with extended algorithms it can be used in multi-class problems. To solve a classification task by a supervised machine learning model like SVM, the task usually involves with training and testing data, which consists of some data instances. Each instance in the training set contains one “target value” (class labels, where class label 1 for positive and class label -1 for negative target value and several “attributes” (features). The goal of a supervised SVM classifier method is to produce a model which predicts target value of the attributes. For each SVM, there are two data set namely, training and testing, where the SVM used the training set to make a classifier model and classify testing data set based on this model with use of their features.

2.2.5 Decision Tree (DT):

DT is a powerful and popular tool for classification and prediction [7]. The attractiveness of DT is due to the fact that in contrast to neural network, it presents rules. Rules can readily be expressed so that human can understand them or even directly use them in a database access language like SQL so that records failing into a particular category may be tree. Decision Tree is a classifier in the form of a tree structure where each node is either a leaf node-indicates the value of the target attributes(class) of expressions, or a decision node that specifies some text to be carried out on a single attribute value with one branch and sub-tree for each possible outcome of the text. It is an inductive approach to acquire knowledge on classification.

2.3. Hybrid Model Approach:

In this approach Rule Based approach and Machine Learning approaches are mixed for more accuracy to identify NERs. Here several combinations are used

2.3.1 HMM approach and Rule Based approach
2.3.2 CRF approach and Rule Based approach
2.3.3 MEMM approach and Rule Based approach
2.2.4 SVM approach and Rule Based approach

A List of Feature set used to identify NERs are

- Context word feature
- word suffix
- word prefix
- Parts of Speech Information (POS)
- Rare word
- first word
- contains digit
- gazetteers lists
- Person-Context
- First Name
- Middle Name
- Last Name
- Location Name
- Month Name
- Day Name
- Length
- Stop words
- Position Orthographic information
- First word
- Digit features
- context lists
- Dynamic NE tag
- Numerical word
- Root Information of word
- Context Word Feature

3. Performance Metrics

Precision (P): Precision is the fraction of the documents retrieved that are relevant to the user's information need.

Precision (P) = correct answers/answers produced

Recall (R): Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

Recall (R) = correct answers/total possible correct answers

F-Measure: The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is

F-Measure = (β2 + 1) PR/ (β2 R + P)

B is the weighting between precision and recall typically β=1. When recall and precision are evenly weighted i.e. β=1, F-measure is called F1 measure. F1 - measure = 2 PR/ (P+R) There is a tradeoff between precision and recall in the performance metric.

4. Observations and Discussions

Now we provide a survey of research done in India looking forward to develop Named Entity Recognition for Indian Languages.

According to the proceedings of IJNLP-08 workshop on NER for South and south east Asian languages which was held in 2008 at IIT Hyderabad had focused on five Indian languages-Hindi, Bengali, Oriya, Telugu and Urdu.

A Recent research work on the Indian Languages is (Sujan Kumar Saha et al.,2008), “A hybrid Approach for Named Entity Recognition in Indian Languages” [3] using Maximum Entropy Markov Model , language dependent rules and Gazetteers have taken into consideration with 12 classes of NER for Hindi, Bengali, Oriya, Telugu and Urdu. For further improvements text from Shakthi standard format is converted into IOB format and tested with More than 5,00,000 of Hindi, 1,60,000 of Bengali 93,000 of Oriya 64,000 of Telugu and 36,000 of Urdu words have been used. The

evaluation has reported F-Score of 65.13% for Hindi, 65.96% for Bengali,44.65% for Oriya, 18.75% for Telugu and 35.47% for Urdu respectively

(Asif Ekbal et al., 2008), Language Independent Named Entity Recognition in Indian languages [4] used the statistical Conditional Random Fields (CRF). The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various named entity (NE) classes. The system uses both the language independent as well as language dependent features. The language independent features are applied to Hindi, Bengali Oriya, Telugu and Urdu and Language dependent features are applied to only Bengali and Hindi. The system has been trained with Bengali(122,467 tokens), Hindi (502,974tokens), Telugu (64,026 tokens), Oriya(93,173 tokens),and Urdu(35,447 tokens) and tested with Bengali(30,505 tokens), Hindi (38,708 tokens), Telugu (6,356tokens), Oriya(24,640 tokens),and Urdu(3,782 tokens) and found the maximal F-measure of 53.46% for Bengali. And very poor F-measure was found for Telugu.

(Praneeth M Shishtla et al, 2008), "A Character n-gram Based Approach for Improved Recall in Indian Language NER" [5] used Conditional Random Fields with Character based n-gram technique on two languages Telugu and Hindi with annotated Telugu corpus containing 45,714 tokens out of which 4709 were named entities, English corpus contained 45,870 tokens out of which 4287 were named entities and Hindi corpus contained 45,380 tokens out of which 3140 were named entities. A total of Nine features were used in training and testing and not used any of the language dependent resources and used POS taggers, Chunkers, morphological analyzers... etc and also included some regular expressions and gazetteer information. Gram n=3 gave better F-measure up to 24.2% for 10k words, 35.38% for 20k words, 44.48% for 30k words and 48.93% for 35k words for Telugu, Gram n=2 gave better F-measure up to 52.92% for 10k words, 65.59% for 20k words, 67.49% for 30k words and 68.46% for 35k words for English and Gram n=4 gave better F-measure up to 40.96% for 10k words, 36.26% for 20k words, 42.36% for 30k words and 45.18% for 35k words for Hindi. The evaluation achieved an over all F-measure of 49.62% for Telugu and 45.07% for Hindi. More number of tested words giving a maximum F-measure.

(P Srikanth and K. Narayana Murthy 2008), [6] "Named Entity Recognition for Telugu" developed a Conditional Random Fields approach with rule based NER System for Telugu trained on a manually tagged data of 13,425 words and tested on a test data set of 6,223words and recorded 92% of F-measure which was manually checked. Named Entity tagged corpus of 72,157 words has been developed using the rule based tagger through bootstrapping and features (length, stop words, affixes, position, POS Orthographic information, suffixes) are applied on three named entity classes person, place and organization. The result obtained by this approach resulted in an impressive F-measure between 80% and 97%.

(Asif Ekbal and Sivaji Bandyopadhyay 2008),[7] " Bengali Named Entity Recognition using Support Vector Machine" used Support vector Machine approach on Bengali with training set of 1,30,000 words with Sixteen Named entity tags using BIE(beginning, intermediate, ending) model for Person, Location, and Organization. This model includes gazetteers with 20,455 person names, 11,668 location names, 963 organization names and 11,554 miscellaneous words and tested on1,50,000 The evaluation has reported good F-measure of 91.8%.

(Vijaya krishna R and Sobha L 2008),[8] "Domain Focused Named Entity Recognition for Tamil Using Conditional Random Fields" developed domain focused Named Entity Recognizer for tourism domain using Conditional Random Fields approach on Tamil language. To improve the performance they have used 106 tag sets for tourism domain and Five feature templates. A 94,000 words corpus is collected in Tamil for tourism domain. Morph analysis, POS tagging, NP chunking and named entity annotation are done manually done on the corpus. This contains about 20,000 named entities and split into two sets. One forms the training data and the other forms the test data. They consist of 80% and 20% of the total data respectively. A total of 4059 named entities are tested for experiment and got F-overall F-measure 80.44%.

Ambiguity in Telugu

Person name Vs Organization name:

Daa. reDDi (Dr. Reddy)
Vs
Daa. reDDi lyaabs(Dr. Reddy labs)

satyaM (Satyam)
Vs
satyaM coMpyuTars (Satyam computers)

Person name Vs Place:

raMgaareDDi (Rangareddy)
Vs
raMgareDDi jillaa (Rangareddy District)

Person name Vs Common nouns:

baMgaaru laksman
(bangaru laxman)
Vs
baMgaaru golusu
(gold chain)

Place Vs Organization:

vijayavaaDa (vijayawada)
Vs
vijayavaaDa tharmal pavar sTashan
(Vijayawada thermal power station)

Appearance in various forms:

telugu dees`aM paarTii (Telugu desam party) , Ti.Di.pi (T.D.P) tee.dee.paa, Ti Di pi. tee dee paa, TiDipi. Teedeepaa.

Table 1: The Named Entity approaches, Training and Testing data on various Indian languages.

Author	*Methods used*	*Indian Language*	*Training data (words)*	*Testing data (words)*
[3]	MEMM	Hindi Bengali Oriya Telugu Urdu	5,00,000 1,60,000 93,000 64,000 36,000	38,704 32,796 26,988 7,076 12,805
[4]	Language independent features	Oriya Telugu Urdu	93,173 64,026 35,447	6,356 24,640 3,782
	Language dependent features	Hindi Bengali	502,974 122,467	38,708 30,505
[5]	Character based n-gram technique	Telugu Hindi	10,000 35,000	45,714 45,380
			60,525	13,425
[6]	CRF	Telugu		
[7]	SVM	Bengali	1,50,000	20,000
[8]	CRF	Tamil	75,200	18,800

Table 2 The Named Entity approaches and F-measures on various Indian languages

Author	*Methods used*	*Indian Language*	*F-measure (%)*
[3]	MEMM	Hindi Bengali Oriya Telugu Urdu	65.13% 65.96% 44.65% 18.75% 35.47%
[4]	Language independent features	Oriya Telugu Urdu	28.71% 47.49% 35.52%
	Language dependent features	Hindi Bengali	33.12% 59.39%
[5]	Character based n-gram technique	Telugu Hindi	48.93% 45.18%
[6]	CRF	Telugu	92%
[7]	SVM	Bengali	91.8%
[8]	CRF	Tamil	80.44%

5. Conclusion

Not much research could be done on Indian languages (especially in Telugu) for reasons of being agglutinative, usage of different possible writing methodologies etc., We conclude that Indian languages are not much researched for NER for various reasons such as agglutinative nature and different kinds of writing styles. Apart from this, there is no concept of Capitalization, difficult Morphology and little availability of annotated corpora. We observed that only statistical approaches for Indian languages may not give good result because of insufficient training data. The training data is in thousands of words only compared to English. BNC contains 100 million annotated corpora available in

English Language. Indian languages are not having readily available such corpora. But only plain corpora is available, that too up to 3 to 10 million Word plain corpora. We observed that rule based approaches may give satisfactory results with sufficient gazetteers list and language independent rules. Language dependent rules are specific for each language. Named entities are open class words, every day new words added to languages and gazetteers list is long. To store all words in gazetteers is a practical difficulty. So gazetteers are needed to divide into finite lists like suffix, prefix context words etc., All Rule based approaches are language dependent. We intend to implement language independent NER system for Indian languages, where Rule based system is not possible. Our conclusion is that development of Hybrid models (gazetteers list, features and statistical methods) may yield improved result for Indian languages.

References

[1] Bh.Krishna Murthy and J.P.L.Gywnn. 1985. *A Grammar of Modern Telugu*. Oxford University Press, Delhi.

[2]CRF++:http://crfpp.sourceforge.net/ *Yet Another CRF toolkit* (accessed on 3 rd may 2010)

[3]Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dantapat,Sudeshna Sarkar and Pabitra Mitra 2008 "A Hybrid Approach for Named Entity Recognition in Indian Languages" Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.

[4] Asif Ekbal, Rajewanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay 2008 "Language Independent Named Entity Recognition in Indian Languages" Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.

[5] Praneeth M Shishtla, Prasad Pingali, and Vasudeva Varma 2008 "ACharacter n-gram Based Approach for Improved Recall in Indian Language NER s" Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.

[6] P Srikanth and Kavi Narayana Murthy 2008 "Named Entity Recognition for Telugu" Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.

[7] Asif Ekbal and Sivaji Bandyopadhyay 2008 " Bengali Named Entity Recognition using Support Vector Machine" Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.

[8] Vijayakrishna R and Sobha L 2008 "Domain Focused Named Entity Recognition for Tamil Using Conditional Random Fields" Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.